TENSOR PRODUCT OF KERNEL MODELS

OMAR DE LA CRUZ C., ALEX BARNETT, HUA TANG AND SUSAN HOLMES

Short Abstract. Kernel methods extend the range of applicability of linear multivariate statistical methods to non-linear settings. Since the eigendecomposition of the kernel matrix is a key step of these methods, it is important to characterize the behavior of the eigenvectors, under different choices of kernel, when the structure of the data is known and is simple. For example, when there is a one-dimensional gradient, the scatterplot of the top two eigenvectors often exhibits the "horseshoe" pattern.

Here we address the characterization of eigenvectors when there is an underlying two-dimensional gradient, by considering the Kronecker product of kernels. However, for this approach to be useful beyond the case of rectangular grids, we use a model-based approach to kernel methods. This approach allows one to qualitatively describe the behavior of the eigenvectors in particular the presence of nodal domains.

Kernel methods are based on extracting global properties of the data from pairwise comparisons between the observation in the study. The underlying intuition is that each observation is represented by and element of a space $\mathcal{X}$ (which we call the *landscape*), and the pairwise comparisons are provided by a kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$; if the observations in a sample are represented by $x_1, \ldots, x_n \in \mathcal{X}$, we obtain a kernel matrix $K = (\mathcal{K}(x_i, x_j))_{ij}$. By Moore–Aronszajn's Theorem [Berlinet and Thomas-Agnan, 2004], if $\mathcal{K}$ is symmetric and of positive type (i.e., K is a positive semidefinite matrix for any $x_1, \ldots, x_n \in \mathcal{X}$ and any $n \geq 1$), then there is a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and a mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that $\langle \phi(x_1), \phi(x_2) \rangle = \mathcal{K}(x_1, x_2)$, for all $x_1, x_2 \in \mathcal{X}$. This way, any linear multivariate statistical method that depends only on inner products of vectors representing the observations, like principal components analysis (PCA), discriminant analysis, and many more, can be applied by simply using K as a Gram matrix of inner products. Amazingly, $\mathcal{H}$ and ϕ do not need to be described; in fact, even $\mathcal{X}$ and the representations of the data points therein can be skipped, as long as we have a method for pairwise comparisons that produces positive semidefinite matrices K. This was observed in [Schölkopf et al., 1998]. When the data is provided as a matrix $X_{n \times p}$ containing measurements of p variables for n observations, so that the observations are naturally represented by n points in $\mathbb{R}^p$, the use of *non-linear kernels* is a very easy way to implement non-linear versions of classical multivariate methods. (The actual Gram matrix XX^T is called the *linear kernel*).

Unfortunately, the great generality of kernel methods can make it difficult to interpret the results. For example, when the data correspond to a set of objects with an underlying one-dimensional characteristic (often called an *ordination*; also a one-dimensional *gradient* or *cline*, as opposed to the presence of clusters), kernel PCA often produces a horseshoe pattern in the scatterplot of the first two components [Diaconis et al., 2008]. A simple examination of the scree plot (eigenvalues plotted in descending order) would suggest that both components contain important information, but it would be a mistake to conclude that there is a two-dimensional structure (a two-dimensional gradient) underlying the data.

One approach to a more systematic understanding of the results of kernel methods is to study what happens when the original data points follow a simple pattern. This is the

approach used in [Diaconis et al., 2008], where the data points were assumed to lie on a one-dimensional, equally spaced grid. This allowed for an explicit description of the eigenvectors of the kernel matrix obtained using the kernel function $\mathcal{K}(x, y) = e^{-|x-y|}$ (a kernel that down-weighs the contribution of comparisons between points far apart, and gives more weight to local information). The question we tackle here is: what is the equivalent of the horseshoe pattern, when the data follow an underlying two-dimensional gradient?

1. Grid case

1.1. **Options.** An obvious approach is to consider a two-dimensional grid. However, there are several natural options to choose from for defining a kernel in this setting. Several of these choices lead to the eigenvectors of the kernel matrix for the two-dimensional grid being Kronecker products of the eigenvectors for the one-dimensional grids:

(1) L^2 (euclidean) distance: The Gaussian radial basis function $\mathcal{K}(\vec{x}, \vec{y}) = e^{-d(\vec{x},\vec{y})^2/2}$ factors:
$$\mathcal{K}_2[(x_1, x_2), (y_1, y_2)] = \mathcal{K}_1(x_1, y_1)\mathcal{K}_1(x_2, y_2),$$
so K_2 is the Kronecker product two copies of K_1.
(2) L^1 (city block) distance: The exponential kernel $\mathcal{K}(\vec{x}, \vec{y}) = e^{-d(\vec{x},\vec{y})}$ factors, as in the case above. Also, this distance is equivalent to the graph distance, with the two-dimensional grid being the cartesian product of two path graphs. So, the adjacency matrix is the Kronecker *sum* of the adjacency matrices of the paths.
(3) L^∞ (max) distance. This is the same as the graph distance, if we add edges joining the diagonally opposite corners of each grid cell. This graph is obtained as the tensor product of two path graphs with loops at each vertex. Then the adjacency matrix is the Kronecker product of the adjacency matrices of the paths.

In case (3), the eigenvectors of the adjacency matrix are the Kronecker products of the eigenvectors the adjacency matrices for the paths, with the eigenvalues being the products of the eigenvalues; surprisingly, in (2) the eigenvectors are the same, but the eigenvalues are *sums* of eigenvalues. Thus, even if the eigenvectors are the same, the order of the corresponding eigenvalues might be different.

1.2. **Patterns.** In the cases above, we have a full description of the eigendecomposition of K_2 based on the eigendecomposition of K_1 (which in turn was characterized in [Diaconis et al., 2008]), by taking the Kronecker product of the eigenvectors. However, due to symmetry, there are many repeated eigenvalues, as is to be expected; the bases for each of those eigenspaces of dimension 2 are not unique, since they can be rotated. Therefore, the eigenvectors produced by a standard numerical algorithm applied to the product kernel are often rotated versions of the Kronecker products of eigenvectors. See Figure 1.2 for a typical example.

When plotted onto the original grid, the obvious pattern is the presence of nodal domains (regions where the entries of the eigenvectors are all positive or all negative). These domains become smaller, and the alternating patterns more complex, as the eigenvalues become smaller.

2. Kernel models

2.1. **Motivation.** A drawback of the rectangular grids considered in the previous section is that they approximate the case where the data points are evenly spread over a rectangular

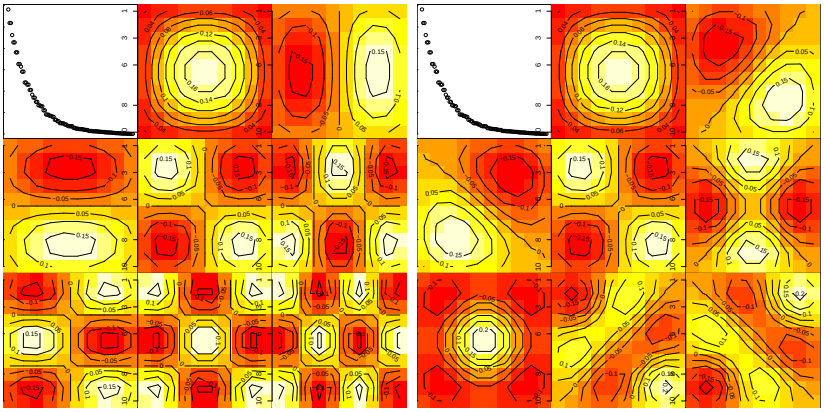

FIGURE 1 Eigenvectors for a two-dimensional grid, using the Gaussian kernel. *Left*: Eigenvalues, and top eight eigenfunctions, obtained as Kronecker products of eigenfunctions of the one-dimensional Gaussian kernel. *Right*: Eigenvalues and eigenfunctions, computed directly from the two-dimensional Gaussian kernel. Some pairs of eigenvectors can be rotated to match the products (the order of the eigenfunctions is also changed).

area; this is not very common in practice, since data points tends to appear in a more rounded blob, with higher concentrations towards the middle.

One way to address the problem of the rectangular shape is to take a model-based approach to kernel methods. The model we describe here is similar to those considered in [Rosasco et al., 2010] and [Smale and Zhou, 2009]. According to this point of view, the structure of the landscape $\mathcal{X}$ is determined by the kernel function $\mathcal{K}$, in the sense that we can use $\mathcal{K}$ to decide how "smooth" a function $f : \mathcal{X} \to \mathbb{R}$ is. Indeed, $\mathcal{K}$ determines a smoothing operator L on a space of functions on $\mathcal{X}$, and its top eigenfunctions are then good candidates for a set of coordinate functions on $\mathcal{X}$. At the same time, the matrix K determines an operator on $\mathbb{R}^n$, which can be seen as a discrete approximation of L. This way, the study of kernel methods can be phrased as "learning the operator L from empirical discrete approximations."

In this view, it becomes obvious that the probability distribution used to draw a sample of points from $\mathcal{X}$ is a key ingredient. Indeed, disregarding the fact that some parts of $\mathcal{X}$ are sampled more densely than others can lead to misinterpretations of the results.

2.2. **The model.** The statistical model includes:

(1) The landscape space $\mathcal{X}$, where the samples come from, together with a notion of smoothness for real-valued functions on $\mathcal{X}$, which is given by the choice of a kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, symmetric and of positive type. These two things come together: in a way, the space is implicit in the kernel, and we know about the space only through the kernel.
(2) A sampling probability measure P on $\mathcal{X}$; the n observations in the study are assumed to have been sampled i.i.d. according to P.

In fact, what one is choosing is a RKHS of functions on $\mathcal{X}$ whose elements will be those functions declared to be smooth on $\mathcal{X}$.

Assuming there is a reference probability distribution Q on $\mathcal{X}$ (for example, uniform), we have two smoothing operators acting on $\mathscr{L}^2(\mathcal{X}, P)$:

$$T_{\mathcal{K}} : f \mapsto \int_{\mathcal{X}} f(y)\mathcal{K}(\cdot, y)\, dQ(y) \quad \text{and} \quad S_{\mathcal{K}} : f \mapsto \int_{\mathcal{X}} f(y)\mathcal{K}(\cdot, y)\, dP(y).$$

Remark: Since we want to learn about $\mathcal{X}$, we are usually more interested in $T_{\mathcal{K}}$; but, unless we also estimate P, and adjust accordingly, we will be estimating $S_{\mathcal{K}}$ instead.

2.3. **Products of models.** We can now consider the product of two models $(\mathcal{X}, \mathcal{K}_{\mathcal{X}}, P_{\mathcal{X}})$, $(\mathcal{Y}, \mathcal{K}_{\mathcal{Y}}, P_{\mathcal{Y}})$ to obtain a model with higher intrinsic dimension: The landscape is taken to be the

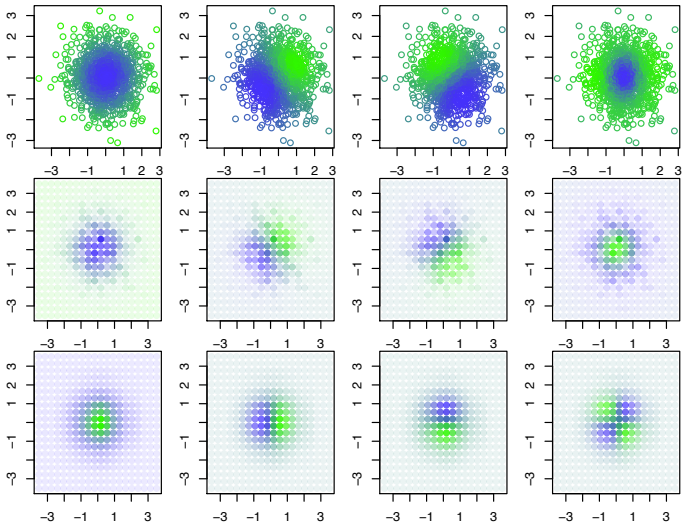

FIGURE 2 Two dimensional case. A sample from a standard bivariate normal was taken; the first row shows the first four eigenfunctions of $S_{\mathcal{K}}$, computed from the points. The second row shows the eigenfunctions computed from the data binned into a square grid; the estimate for the density function is taken from the number of points in each bin. The last row contains the first four eigenfunctions obtained by the product method; the densities are the product of the estimated marginal densities, and the eigenfunctions are the Kronecker product of the one-dimensional eigenfunctions. The transparency level corresponds to lower density, and the colors to the eigenfunctions.

cartesian product $\mathcal{X} \times \mathcal{Y}$, the kernel the Kronecker product $\mathcal{K}_{\mathcal{X}} \otimes \mathcal{K}_{\mathcal{Y}}$, while the sampling probability distribution is obtained as the product measure $P_{\mathcal{X}} \times P_{\mathcal{Y}}$ (thus implying the assumption of independence of the sampling probabilities).

The landscape is still "rectangular," since it is a cartesian product, but the observations come from a probability distribution that is not necessarily uniform on this rectangle (for example, it could be a bivariate normal distribution, if the sampling probability distributions on the factor models were univariate normal).

This allows us to characterize what we can expect to find when there is a two-dimensional gradient structure in the data. This includes the existence of nodal domains: contiguous regions of $\mathcal{X}$ where the eigenvectors have the same sign. These domains can be observed in practice (when $\mathcal{X}$ is not known) by, e.g., plotting a scatterplot of the top two eigenvectors and coloring the points according to the 3rd (or lower) eigenvectors. If well defined regions of similar color can be seen, this is the equivalent of the horseshoe pattern in this setting. Thus, one can learn about the underlying dimension: if nodal domains are observed, this suggests the existence of a two-dimensional gradient; if nodal domains are not apparent in this setting, then the 3rd eigenvector likely represents an extra spatial dimension. An advantage is that each of the factors can be estimated more accurately or stably, by estimating on the margins. Then the product model is estimated in a more stable way from these marginal models (this, however, depends on the assumption of independence of the sampling probability distributions).

REFERENCES

[Berlinet and Thomas-Agnan, 2004] Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA. With a preface by Persi Diaconis.

[Diaconis et al., 2008] Diaconis, P., Goel, S., and Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *Ann. Appl. Stat.*, 2(3):777–807.

[Rosasco et al., 2010] Rosasco, L., Belkin, M., and Vito, E. D. (2010). On learning with integral operators. *J. Mach. Learn. Res.*, 11:905–934.

[Schölkopf et al., 1998] Schölkopf, B., Smola, A., and Muller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

[Smale and Zhou, 2009] Smale, S. and Zhou, D.-X. (2009). Geometry on probability spaces. *Constr. Approx.*, 30(3):311–323.